Cover-based combinatorial bounds on probability of overfitting

A. Frey, I. Tolstikhin

4 сентября 2013 г.

For the last 40 years the problem of accurate overfitting bounds remains an active area of research within Statistical Learning Theory [1, 2]. For some special cases exact bounds were for the first time derived in combinatorial approach [3, 5]. Further work [4] derives general but less accurate combinatorial bounds based on the principle of protective and prohibitive sets. Those bounds were successfully applied to the family of threshold conjunction rules, and lead to a better quality of logical classification in practical machine learning tasks. However, those bounds were only applicable to connected sets of classifiers of a small cardinality.

In this work we propose better overfitting bounds which mitigate some of limitations in combinatorial approach. New bounds are based on a specific cover of the original set of classifieres such that each element of the cover have an exact overfitting bound.

Let $\mathbb{X} = \big(x_1, \ldots, x_L\big)$ be a finite instance space and A be a set of classifiers. By $I\colon A \times \mathbb{X} \to \{0, 1\}$ denote a binary loss function such that $I(a, x) = 1$ if classifier $a \in A$ produces an error on object $x \in \mathbb{X}$. A binary vector $(a_i) = (I(a, x_i))_{i=1}^{L}$ is called an *error vector* of classifier $a \in A$. For an arbitrary $U \subset \mathbb{X}$ and $a \in A$ let $n(a, U) = \sum\limits_{x_i \in U} I(a, x_i)$ denote the *number of errors*, and let $\nu(a, U) = n(a, U)/|U|$ denote the *error rate* of classifier a.

By $[\mathbb{X}]^\ell$ denote the set of all $\binom{L}{\ell} = \frac{L!}{\ell!(L-\ell)!}$ subsets $X \subset \mathbb{X}$ of a fixed size ℓ. Denote $k = L - \ell$ and $\bar{X} = \mathbb{X} \backslash X$. A set $X \in [\mathbb{X}]^\ell$ is said to be a *train sample*, and the corresponding $\bar{X} = \mathbb{X} \backslash X$ is said to be a *test sample*. A

learning algorithm is a mapping of the form $\mu\colon [\mathbb{X}]^\ell \to A$, which takes each train sample $X \in [\mathbb{X}]^\ell$ to classifier $\mu X \in A$. A classifier μX is said to be *overfitted* if the deviation between the test and train error rates $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ exceeds a given threshold $\varepsilon > 0$. Then the *probability of overfitting* $Q_\varepsilon(\mu)$ is defined as the fraction of $X \in [\mathbb{X}]^\ell$ such that μX is overfitted:

$$Q_\varepsilon(\mu) = \mathsf{P}[\delta(\mu X, X) \geqslant \varepsilon], \tag{1}$$

where $\mathsf{P}[\varphi] \equiv \frac{1}{C_L^\ell} \sum\limits_{X \in [\mathbb{X}]^\ell} \varphi(X)$, and φ is an arbitrary predicate on $[\mathbb{X}]^\ell$. Here and in the sequel $[\mathit{true}] = 1$, $[\mathit{false}] = 0$.

Consider a disjoint partitioning of a set of classifeirs A:

$$A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t \tag{2}$$

such that for every A_i we have explicit conditions for $\mu X \in A_i$.

Hypothesis 1. *Let A be a set of classifiers, and $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$ be a disjoint partitioning of A. Suppose for all $i = 1, \ldots, t$ there exist two disjoint sets $X_i \subset \mathbb{X}$ and $X'_i \subset \mathbb{X}$ such that for all $X \in [\mathbb{X}]^\ell$*

$$\mu X \in A_i \Rightarrow (X_i \subset X) \text{ and } (X'_i \subset \bar{X}).$$

In addition, suppose that all classifiers $a \in A_i$ have zero errors on X_i, and have an error on each object $x \in X'_i$.

The set X_i is called *protective* for A_i, the set X'_i is called *prohibitive* for A_i. Hypothesis 1 imply that $\mu X \in A_i$ require all objects from the protective set to belong to the train sample X, and all objects from the prohibitive set to belong to the test sample $\bar{X}$. The set of remaining objects $Y_i \equiv \mathbb{X} \backslash X_i \backslash X'_i$ is called *neutral* for A_i.

For all $i = 1, \ldots, t$ denote $L_i = L - |X_i| - |X'_i|$, $\ell_i = \ell - |X_i|$, $k_i = k - |X'_i|$. By $Q_\varepsilon(A_i)$ denote the upper bound on probability of overfitting for an arbitrary method $\mu\colon [Y_i]^{\ell_i} \to A_i$:

$$Q_\varepsilon(A_i) = \frac{1}{C_{L_i}^{\ell_i}} \sum_{Y \in [Y_i]^{\ell_i}} [\max_{a \in A_i} \delta(a, Y) \geqslant \varepsilon], \tag{3}$$

where $[Y_i]^{\ell_i}$ denotes the set of all samples $Y \subset Y_i$ of size ℓ_i.

Theorem 1 (Splitting and similarity bound). *Consider a partitioning $A = A_1 \sqcup A_2 \sqcup \cdots \sqcup A_t$. Suppose that for every A_i there exist a number m_i such that all classifiers $a \in A_i$ have an equal number of errors $n(a, \mathbb{X}) = m_i$. Then under the assumptions of hypothesis 1 we have*

$$Q_\varepsilon(\mu) \leqslant \sum_{i=1}^{t} P_i \, Q_{\varepsilon_i}(A_i), \tag{4}$$

where $P_i = \frac{C_{L_i}^{\ell_i}}{C_L^\ell}$, $\varepsilon_i = \frac{L_i}{\ell_i k_i}\frac{\ell\, k}{L}\varepsilon + \big(1 - \frac{\ell\, L_i}{L \ell_i}\big)\frac{m_i}{k_i} - \frac{|X_i'|}{k_i}$, and $Q_\varepsilon(A_i)$ is the probability of overfitting (3) *on the set of neutral objects Y_i.*

Practical usage of bound (4) require a generic method for constructing protective and prohibitive subsets for every A_i, $i = 1, \ldots, t$. This in turn require a specific learning method.

Definition 1. *A learning algorithm μ is said to be a pessimistic empirical risk minimisation (or, for short, pessimistic ERM), if $\forall X \in [\mathbb{X}]^\ell$ it satisfies $\mu X \in \operatorname*{Argmax}_{a \in A(X)} n(a, \mathbb{X})$, where $A(X) \equiv \operatorname*{Argmin}_{a \in A} n(a, X)$.*

A *partial order "$\leqslant$" on classifiers* is defined as an element-wise comparison of their error vectors: $a \leqslant b \Leftrightarrow I(a, x_i) \leqslant I(b, x_i), \forall x \in \mathbb{X}$. Let $\rho(a, b) = \sum_{i=1}^{L} |a_i - b_i|$ denote the Hamming distance between the error vectors of classifiers a and b. A pair of classifieres (a, b) is said to be *connected* if $\rho(a, b) = 1$. A *precedence* relation $a \prec b$ is defined as $\big(a \leqslant b\big) \wedge \big(\rho(a, b) = 1\big)$. Then the protective and prohibitive sets for a single classifier $a \in A$ are defined in the [5] as follows:

$$\begin{aligned} X_a &= \{x \in X \colon \exists b \in A \colon a \prec b, I(a, x) < I(b, x)\}, \\ X_a' &= \{x \in X \colon \exists b \in A \colon b < a, I(b, x) < I(a, x)\}. \end{aligned} \tag{5}$$

Lemma 2. *Suppose μ is a pessimistic ERM. Consider the following sets:*

$$X_i = \bigcap_{a \in A_i} X_a, \quad X_i' = \bigcap_{a \in A_i} X_a', \tag{6}$$

where X_a and X_a' are defined by (5)*. Then the set X_i is a protective set for A_i, and the set X_i' is a protective set for A_i in terms of hypothesis 1.*

Lemma 2 enables us to calculate all parts of bound (4), except for $Q_{\varepsilon_i}(A_i)$. To handle this we extend each subset A_i to $B_i \supseteq A_i$, and replace $Q_{\varepsilon_i}(A_i)$ with $Q_{\varepsilon_i}(B_i)$.

Theorem 3. *Let A_i and B_i be two sets of classifiers such that $A_i \subset B_i$. Suppose all $a \in B_i$ have an equal number of errors on $\mathbb{X}$. Then*

$$Q_\varepsilon(A_i) \leqslant Q_\varepsilon(B_i). \tag{7}$$

Theorem 3 implies that the sets B_i have a simple structure and support fast computation of $Q_\varepsilon(B_i)$. The following two definitions provide such examples of B_i.

Definition 2. *Suppose classifier $a_0 \in A$ has m errors on $\mathbb{X}$, r be a natural number, $r < m$. Then the set $B_r^m(a_0)$ is defined as the following subset of $\{0,1\}^L$:*

$$B_r^m(a_0) = \{a \in \{0,1\}^L \colon \rho(a,a_0) \leqslant r \text{ and } n(a,\mathbb{X}) = m\}.$$

Definition 3. *Let $\mathbb{X} = X_0 \sqcup X_1 \sqcup X_r$ be a partitioning of the instance space $\mathbb{X}$. Let $|X_r| = r$, $|X_1| = m$, and ρ is a number such that $\rho \leqslant r$. Then a set $B_{r,\rho}^m$ is defined as a subset of $\{0,1\}^L$ such that*

- *$B_{r,\rho}^m$ consists of all classifiers with ρ errors on X_r,*
- *classifiers from $B_{r,\rho}^m$ have no errors on X_0,*
- *classifiers from $B_{r,\rho}^m$ have an error on each $x \in X_1$.*

Inference of explicit formulas for $Q_\varepsilon(B_r^m(a_0))$ and $Q_\varepsilon(B_{r,\rho}^m)$ utilize intrinsic symmetries of the sets $B_r^m(a_0)$ and $B_{r,\rho}^m$, and is based on randomized learning algorithm [6, 7]. These formulas are applicable for deterministic ERM because in both sets all classifiers share the same number of errors on $\mathbb{X}$.

Randomized ERM [6] is a learning algorithm that for a given $X \in [\mathbb{X}]^\ell$ selects a random classifier from the set $A(X) \equiv \operatorname*{Argmin}_{a \in A} n(a,X)$. The definition of the probability of overfitting (1) is adjusted as follows:

$$Q_\varepsilon(A) = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell} \frac{1}{|A(X)|} \sum_{a \in A(X)} \big[\delta(a,X) \geqslant \varepsilon\big]. \tag{8}$$

Let $S_L = \{\pi\colon \mathbb{X} \to \mathbb{X}\}$ be a symmetric group, which acts on the set $\mathbb{X}$ by permutation of objects. Any $\pi \in S_L$ also acts on a classifier $a \in A$ by permutation of the coordinates of the corresponding error vector: $(\pi a)(x_i) = a(\pi^{-1}x_i)$. For any $X \in [\mathbb{X}]^\ell$ and any set $A \subset \{0,1\}^L$ the actions πX and πA are defined as follows: $\pi X = \{\pi x\colon x \in X\}$, $\pi A = \{\pi a\colon a \in A\}$.

Definition 4. *The symmetry group* $\mathrm{Sym}(A)$ *for a set of classifiers* $A \subset \{0,1\}^L$ *is defined as the stationary subgroup of* S_L*:*

$$\mathrm{Sym}(A) = \{\pi \in S_L\colon \pi A = A\}.$$

Denote by $\Omega([\mathbb{X}]^\ell)$ the set of all orbits of A under the action of group $\mathrm{Sym}(A)$ on $[\mathbb{X}]^\ell$. Let X_τ be an arbitrary element from an orbit $\tau \in \Omega([\mathbb{X}]^\ell)$.

Theorem 4. *Let* A *be a set of classifiers, and* $\mathrm{Sym}(A)$ *be its symmetry group. Then the probability of overfitting* (8) *can be written as follows:*

$$Q_\varepsilon(A) = \sum_{\tau \in \Omega([\mathbb{X}]^\ell)} \frac{|\tau|}{|A(X_\tau)|} \sum_{a \in A(X_\tau)} [\delta(a, X_\tau) \geqslant \varepsilon]. \tag{9}$$

This theorem yields the following expressions for $Q_\varepsilon(B_r^m(a_0))$ and $Q_\varepsilon(B_{r,\rho}^m)$.

Theorem 5. *Suppose* $B_r^m(a_0)$ *is a central slice of classifieres (Definition 2) such that* $r \leqslant 2m$ *and* $n(a_0, \mathbb{X}) = m$*. Then*

$$Q_\varepsilon(B_r^m(a_0)) = H_L^{\ell,m}\left(s(\varepsilon) + \left\lfloor r/2 \right\rfloor\right) \cdot [m \geqslant \varepsilon k], \tag{10}$$

where $s(\varepsilon) = \frac{\ell}{L}(m - \varepsilon k)$, $H_L^{\ell,m}(s) = \sum\limits_{t=0}^{\lfloor s \rfloor} C_m^t C_{L-m}^{\ell-t} / C_L^\ell$ *is the function of hypergeometric distribution [3].*

Theorem 6. *Suppose* $B_{r,\rho}^m$ *is a heap of classifiers (Definition 3). Then*

$$Q_\varepsilon(B_{r,\rho}^m) = \frac{1}{C_L^\ell} \sum_{i=0}^{\min(m,\ell)} \sum_{j=0}^{\min(r,\ell-i)} C_m^i C_r^j C_{L-m-r}^{\ell-i-j} \Big[\frac{m+\rho-t}{k} - \frac{t}{\ell} \geqslant \varepsilon\Big], \tag{11}$$

where $t = i + \max(0, \rho - r - j)$.

Let us describe the final computation scheme for our new bound on the probability of overfitting for an arbitrary set A. We start with an arbitrary partitioning of the set of classifier A into subsets (2). For each subset we build the protective and prohibitive subsets according to (6). Then we embed each subset A_i into a larger superset B_i with know formula for the probability of overfitting (for example, (10) or (11)). The final bound is given by (4).

We evaluated the new bound on 11 datasets from UCI repository, and compare the actual test error rates with our prediction. The prediction of our new bound exceed the actual test error rate by 5 to 50 % (depending on a dataset). This is sharper than the prediction of the splitting-connectivity bound [5], which in our experiments exceeded the actual test error rate by 17 to 63 %. Both combinatorial bounds are much sharper than the latest PAC-Bayesian bounds [8], which exceeded the actual test error rate from 2.5 to 10 times.

This work was supported by the Russian Foundation for Basic Research (project no. 11-07-00480, no. 12-07-33099-mol-a-ved) and by the program "Algebraic and Combinatorial Methods in Mathematical Cybernetics and New Generation Information Systems" of the Department of Mathematical Sciences of the Russian Academy of Sciences.

Список литературы

[1] Vapnik V. N., Chervonenkis A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2), 264–280.

[2] Boucheron S., Bousquet O., Lugosi G. (2005) Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9(1), 323–375.

[3] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[4] Vorontsov K. V., Ivahnenko A. A. (2011) Tight combinatorial generalization bounds for threshold conjunction rules. *4-th Int'l Conf. on Pattern Recognition and Machine Intelligence (PReMI'11).* Lecture Notes in Computer Science, Springer-Verlag, 66–73.

[5] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, No. 3. — Pp. 269–285.

[6] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов и рандомизированных методов обучения // Pattern Recognition and Image Analysis. — 2010.

[7] *Толстихин И. О.* Вероятность переобучения плотных и разреженных семейств алгоритмов // Междунар. конф. ИОИ-8 — М.: МАКС Пресс, 2010. — С. 83–86.

[8] Jin C., Wang L. (2012) Dimensionality Dependent PAC-Bayes Margin Bound. *In Advances in Neural Information Processing Systems*, 25, 1043–1051.

УДК 519.7:004.855.5

Перевод названия, имени и фамилии авторов

Cover-based combinatorial bounds on probability of overfitting

Alexander Frey, Ilya Tolstikhin

Данные об авторах

1. Фрей Александр Ильич

Тел.: +7(903)175-80-77;

Email: sashafrey@gmail.com;

Московский Физико-Технический Институт (Государственный Университет)

2. Толстихин Илья Олегович

Тел.: +7(916)136-49-78;

Email: iliya.tolstikhin@gmail.com;

Учреждение Российской академии наук Вычислительный центр им. А. А. Дородницына РАН.

Abstract

Cover-based combinatorial bounds on probability of overfitting

A. Frey, I. Tolstikhin

The paper improves existing combinatorial bounds on probability of overfitting. A new bound is based on partitioning of a set of classifiers into non-overlapping clusters, and then embedding each cluster into a superset with known exact formula for the probability of overfitting. Such approach makes the bound sharper because it accounts for similarities between classifiers within each cluster.